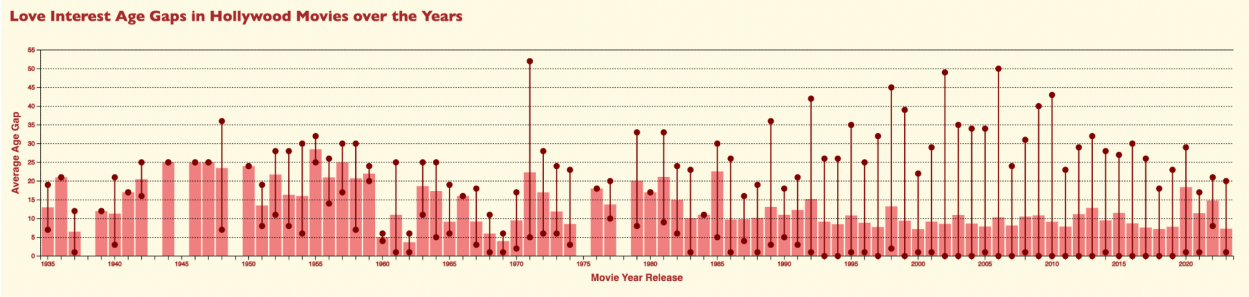
Adeniyi Fagbewesa, Elise Song, Ka-Hyun Nam, Morgan Suh
Info 3300

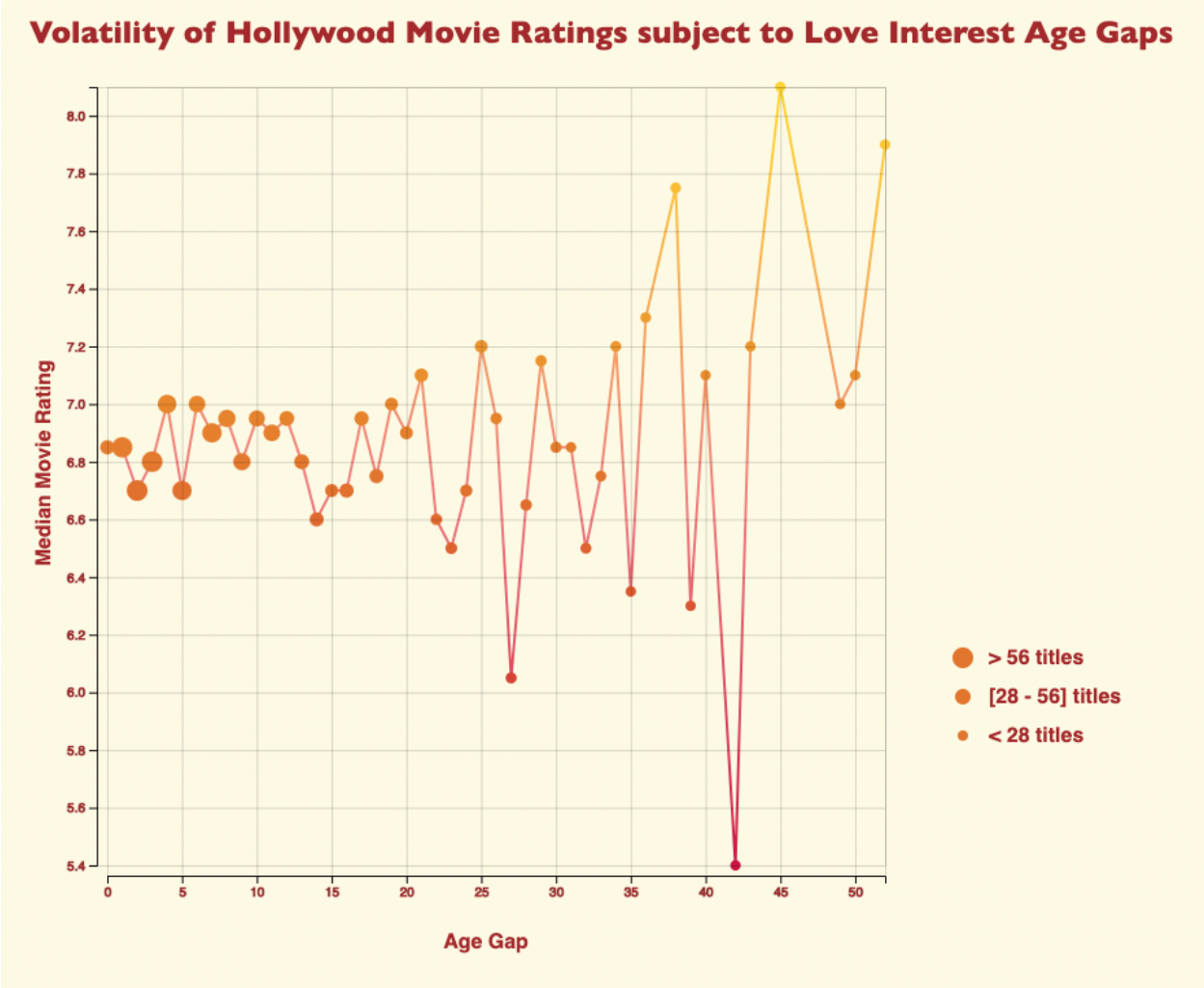
Project 1: Love Interest Age Gaps in Hollywood Movies

Data Visualization:

Plot 1



Plot 2



Data Description:

The movie relationships and actor age dataset was sourced from a kaggle dataset (https://www.kaggle.com/datasets/sujaykapadnis/hollywood-age-gap/) and the movie rating data was compiled from the OMDb API (https://www.omdbapi.com/). The kaggle dataset was reorganized by grouping relationships in the same movie under the same data item in the final visualization dataset. There were fields available in the kaggle dataset (for example, the gender of each actor in the relationship) that were not selected due to being irrelevant to our general thesis. The ratings data was organized using a Python script to request each movie's information (iterating over the movie title of each movie in the kaggle dataset) from the API and writing the Imdb rating field from the response to the rating.csv file. Some of the movies from the kaggle dataset were unavailable on the OMDb API, in which case the pre-processing filtered the data without associated ratings data from the final dataset used for the visualization.

Design Rationale:

For Plot 1, the marks are the bars and the vertical lines along with the circles. The channels are the horizontal and vertical aligned positions of the circles, bars, and vertical lines. In Plot 1, the bars show the average age difference between actors of movie love interests each year with the vertical lines along with the circles showing the lower and upper range of age gaps of that year. The bars are filled with a coral color to make them distinct and the lines are colored in maroon so they are also visible in front of the bar chart. For the x-axis it is scaled using the d3.scaleBand() to create a categorical scale for movie release years so that it is evenly distributed. The y-axis is linearly scaled however it is also padded to make sure the maximum age range is visible and comprehensible. One decision that was made was to include the range of age gaps to give more information about the variability in age differences for each year.

In Plot 2, both circles and trend lines serve as data marks. Circles represent individual data points, with their position, size, and color acting as visual channels. While trend lines utilize the exact visual channels, they additionally capture the progression from point to point using slope as a channel. Both axes employ linear scales for clarity, given the even distribution of median age and ratings. The y-axis fits the median movie ratings range of [5.4, 8.5] to aid visualization and reduce the amount of empty space. One thing to acknowledge is the convergence to the mean in our data – that the results are a bit skewed since there is a higher density of movies with smaller age gaps, shown by the radius of the points. We chose to use the median age gap instead of average for this reason. The size of the circles are positively correlated to the density of movies with that data's age gap; this was done to show more information and to avoid being misleading. Finally, a sequential color scale is applied to circles, while lines use a linear gradient, adhering to our divergent color scheme: red signifies lower-rated movies and yellow indicates higher ratings. We added a key to indicate the radius size of the points after feedback from the demo day. It was not immediately obvious that the radius size of the points were different so the key highlights this and also contextualizes the radius size by the amount of movies for each age gap.

Visualization Story:

Plot 1: Love Interest Age Gaps in Hollywood Movies over the Years

This graph plots the age gap between the actors of love interests in Hollywood movies released from 1935 to 2023. The bars indicate the average age gap of love interests for each year, and the two points connected by a line illustrate the range of the age gaps found in that year. In terms of just the bars, we see the following trends:

- Until 1960, the average age gap was less than 30 years.
- From 1960 to 1970, the average age gap was relatively low, with the majority of age gaps being less than 10 years.
- From 1970 to 1985, the average age gap went back up to less than 25 years.
- From 1985 to the present, the average gap does not go above 15 years, except for an outlier in 2020.

In terms of the ranges, the results are a bit skewed since more movies were produced as the time progresses. However in 1970, the range points show an age gap around 50 years, the highest age gap in the plot. Before that, age gaps of 0 are less common. After that, the age gaps range from 0 to 50 years. Oddly, from 1970 to the mid 2010s, the ranges got larger, and from the mid 2010s, the ranges became smaller.

We wanted the viewers to find trends in the love interest age gaps over time and consider why those trends exist. Many movies have love interests with a young woman and a middle-aged/older man which could be a decision made by the director or a reflection of the times. The age gaps could indicate what is seen as attractive during that time period, and what is deemed socially acceptable. In 1975, the blockbuster era of Hollywood began, and from then, the average age gap tends to be between 5-10 years. However, the age gap ranges indicate that large age gaps still persist.

Plot 2: Volatility of Hollywood Movie Ratings subject to Love Interest Age Gaps

This graph plots the median movie rating per age gap between actors playing love interests. Each point's size indicates the number of movies that were averaged per age gap. The lines between each point show the displacement in ratings between each age gap. For the most part, until an age gap of 20, the average ratings range from 6.6-7. After that, the median movie ratings spike up and down. When coming up with the graph concept, we were unsure if the age gaps would have an effect on the movie rating since there are many factors that make a movie good or bad. Evidently, a larger age gap does not guarantee a high or low rating, but seems to generate a large range of differing opinions. As the age gap increases, the plot becomes more volatile.

We want the viewer to consider if they think about age gaps as a factor of how much they liked the movie, and if the ages of actors affect the perceived chemistry between love interests. Large age gaps can cause feelings of discomfort, disgust, etc., but the plot of the movie may trump that and still achieve a good rating.

Team Contributions:

- Ka-Hyun: Data sourcing (Python script for requesting data from ratings API and writing relevant data from response into csv file), data preprocessing for merging ratings dataset and kaggle movie dataset into final visualization dataset (total 5 hours).
- Elise: Styled plots with cohesive theme (2 hours), formatted write ups (1 hour), wrote visualization story (2 hours), presented visual at demo day (1 hour).
- Morgan: Created the average age differences for love interests for each year bar chart (2 hours), created the age gap range function to create the vertical lines (2 hours), changed formatting to fit both range and bar chart and also style graph (1 hour).
- Adeniyi: Created the movie ratings vs age gaps plot (4 hours, with revisions), enabled stylization properties on the plot (1 hour).